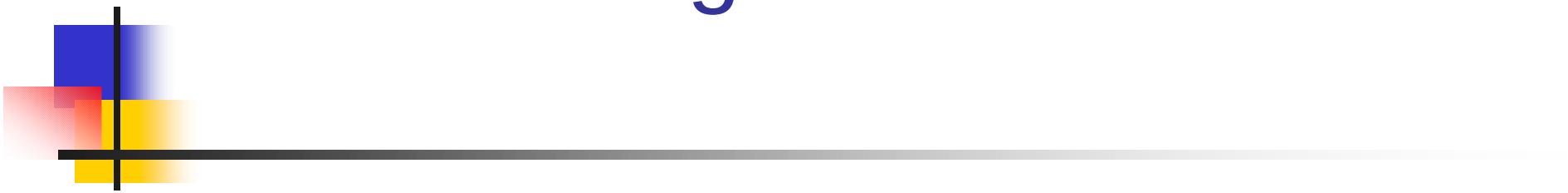
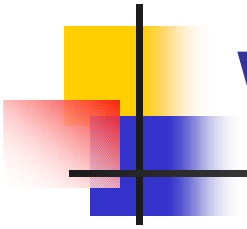
How should we publish data analyses in the web age?



Todd L. Graves

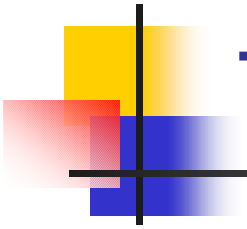
Los Alamos National Laboratory

April 8 2000



What this talk is about

- How does the web allow papers and journals to be different than they have been before? Data analyses have some special characteristics for which the web could be particularly helpful
- NOT about how to put existing papers or journals on the web
- Think of papers about data analyses, not about statistical methodology

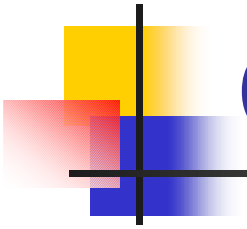


Thanks to

- Visualization group at Bell Labs (Naperville, IL)
- National Institute of Statistical Sciences

Target audience for new journal types

- People who read papers and think:
  - I want those data
  - I want that tool
  - I want to try out my own tool on this problem
  - I want to add my own data to this problem
  - I bet those morons screwed up the analysis
  - Where are all the other data analyses about this problem?
  - I just want to add my own two cents



Goals

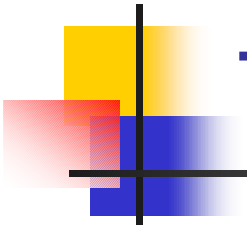
- You're reading a paper. You find the authors' discussion of the data analysis less than totally convincing. You can:
  - Play through authors' analysis step by step
  - Make (small?) changes and examine the result
  - Try out your own model/algorithm/tool
  - Add your own data to the author's
  - After completing your analysis, press a button to submit your comments to the journal and request refereeing

Reading data analyses on the web

- “Live Documents”: see papers by Mockus and others
- Interactive visualization and analysis applets
- Window containing steps of author’s analysis, which reader can play through
- ... and try out modifications. Are the author’s conclusions still valid if ...?

Introducing new (or old) data

- Popular press reports latest research findings as if no previous work exists
- If result is controversial and relatively well-studied, there will often be other relevant data, other analysis approaches, and other willing analysts
- Other data that bear on the problem should ideally be possible to add to the data set with a mouse click



Trying out your own tool

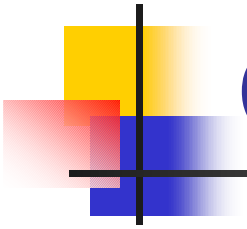
- Must write own code in Java, implementing the appropriate interface
  - Generate “AnalysisAction” objects, handle events generated by user interface, send notification to other components
- Security issues: subscribers to journal receive an encryption key

Submitting your own analysis

- After coming to your own conclusions, press a button to submit your own analysis.
- Security issues
- Presumably data analytic journals will not attract too many garbage submissions if subscription is required

Organization of a new journal

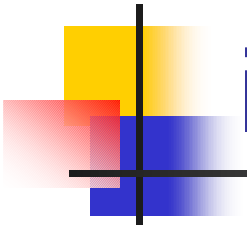
- Papers about the same topic (e.g. at what ages should all women have mammograms? Do particulates of a certain size cause death?) should be kept together, rather than papers published at the same time
- In this way, all data relevant to a given policy question are close at hand



Organization, cont.

- Journal includes a page containing links to “all” submissions and makes no claims about their quality
- But the front page links only to “refereed” submissions

Other applications for these ideas



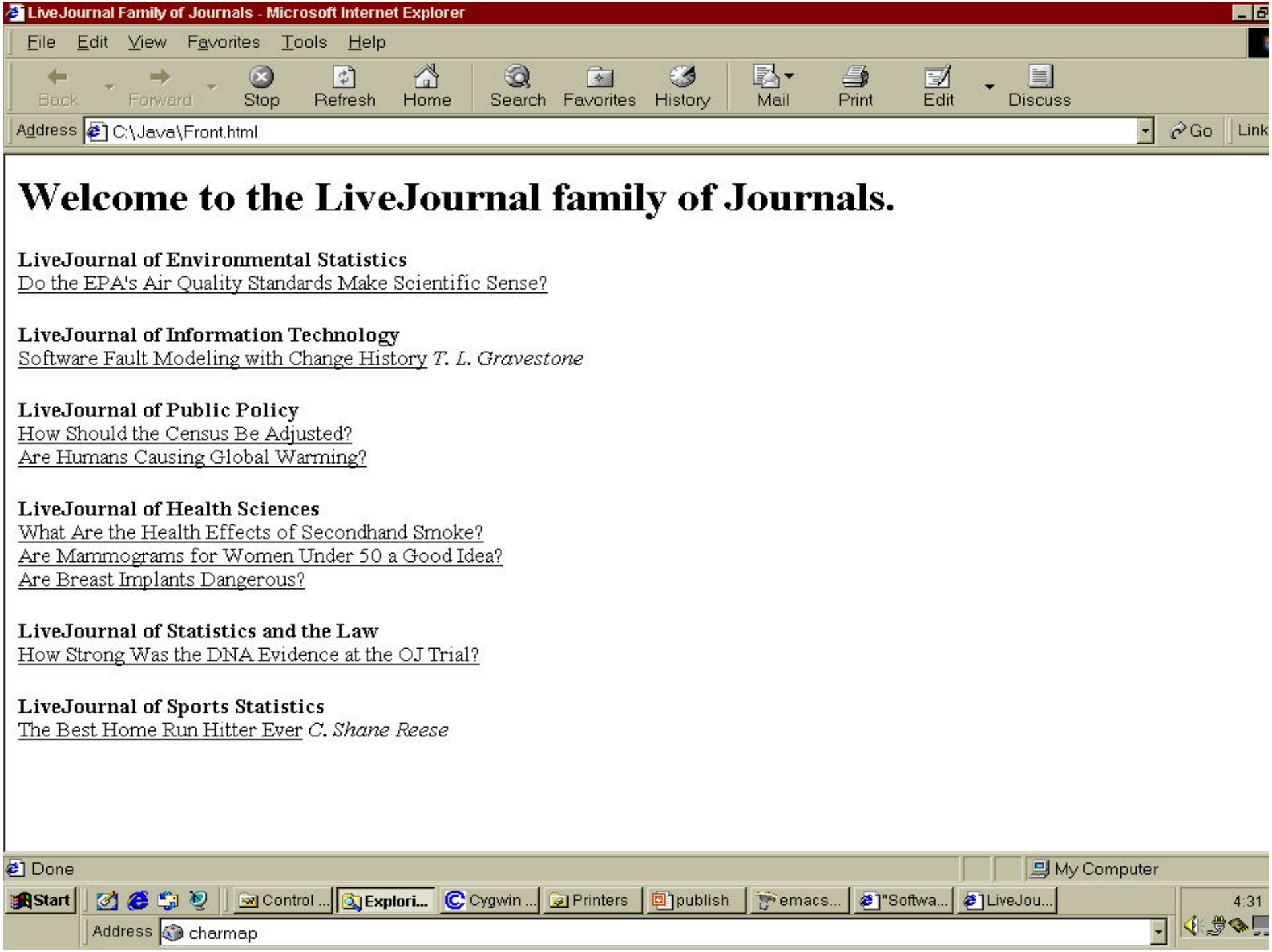
- Overcoming hostility to Bayesian analyses of controversial topics, if readers can pick own priors (and MCMCs run instantly...)
- Notebooks for individual researchers' analyses
- Collaborative research
- Education

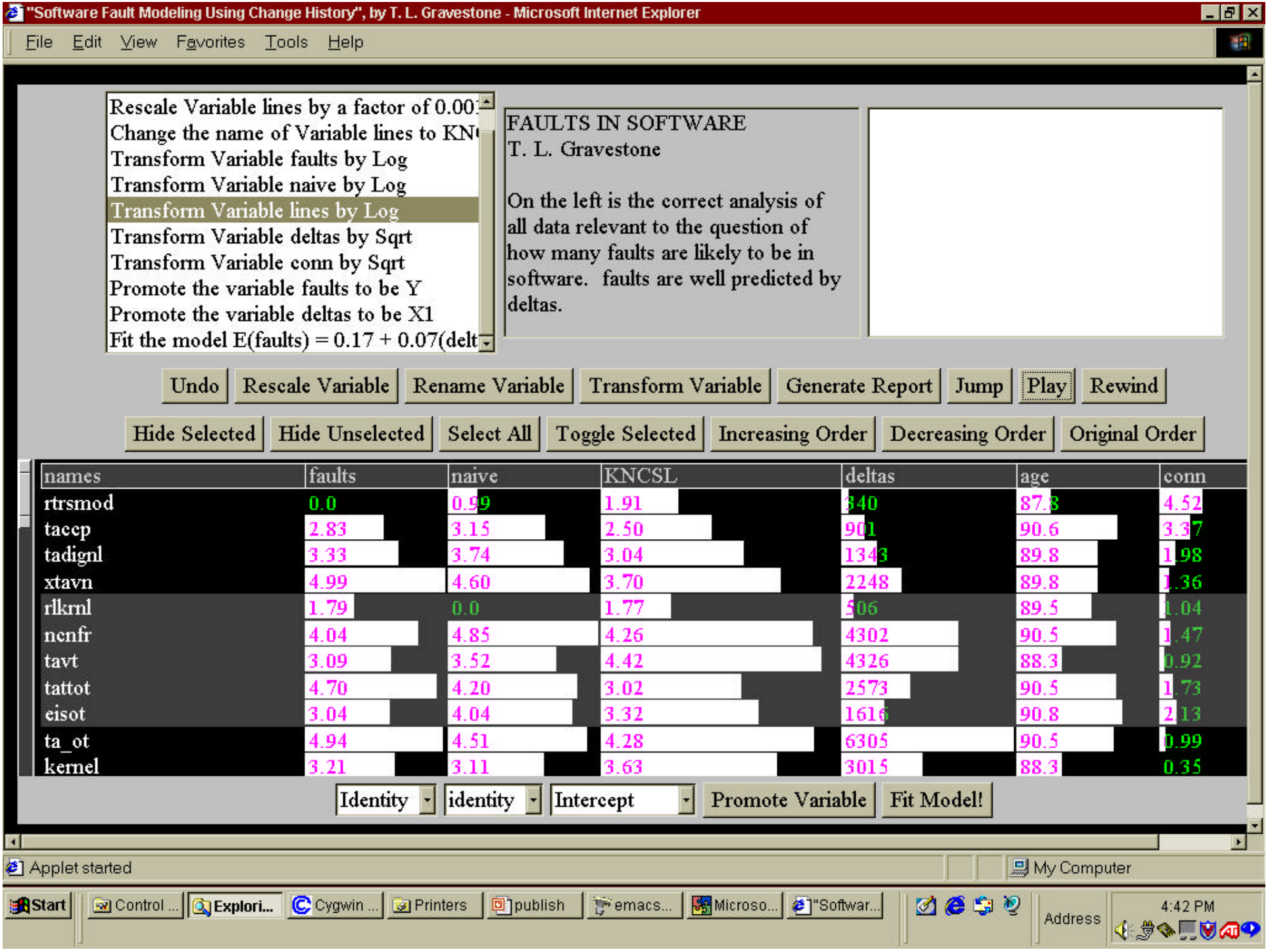
Among the many obstacles

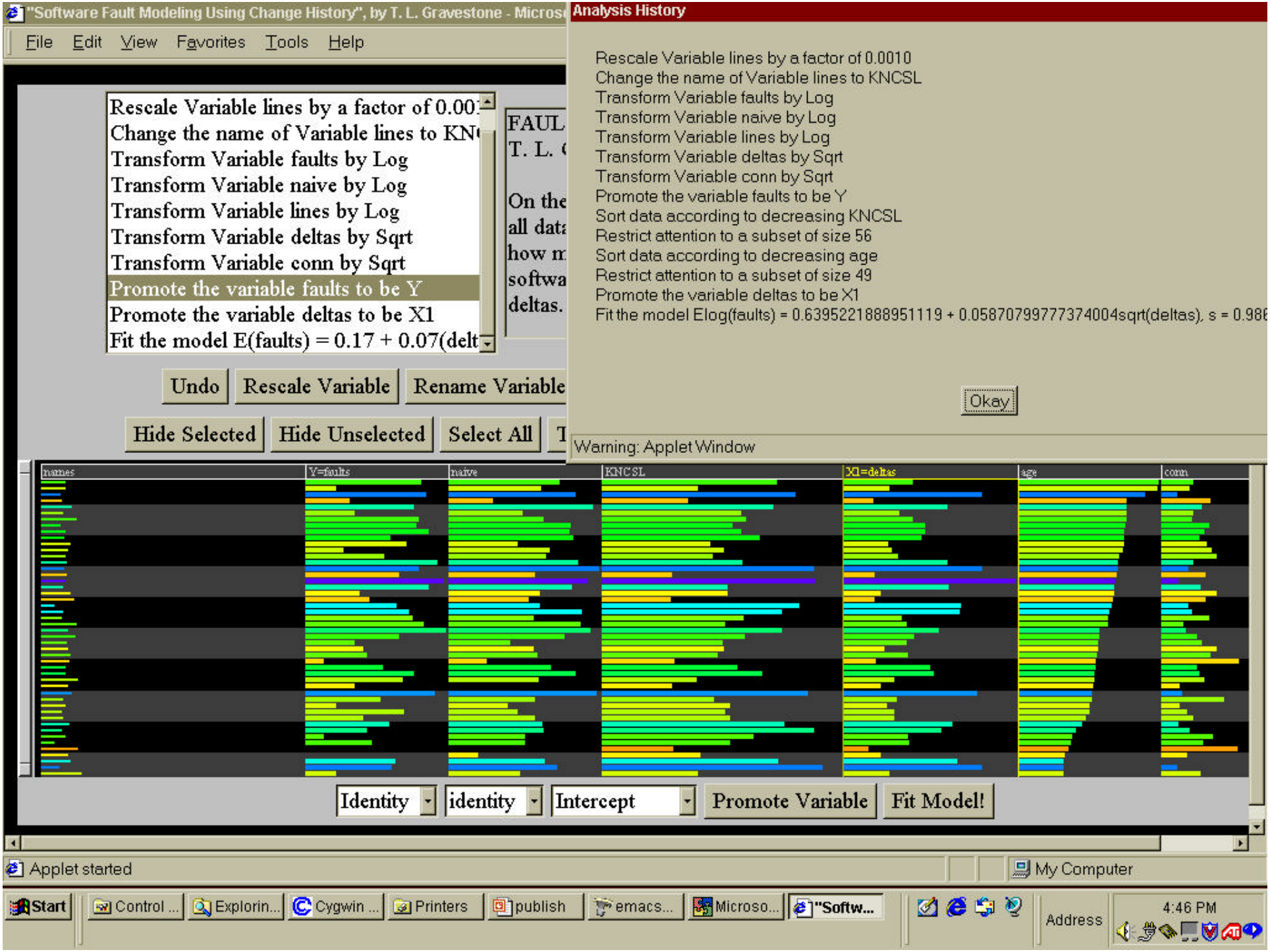
- Setting up infrastructure: funding? Army of volunteer developers?
- Greater difficulty of writing live documents
- Porting software to conform to interactive journal requirements
- Researchers often don’t want to “give up” their data
- Security issues
- Unfamiliar publication metaphor: can such a journal ever earn respect?

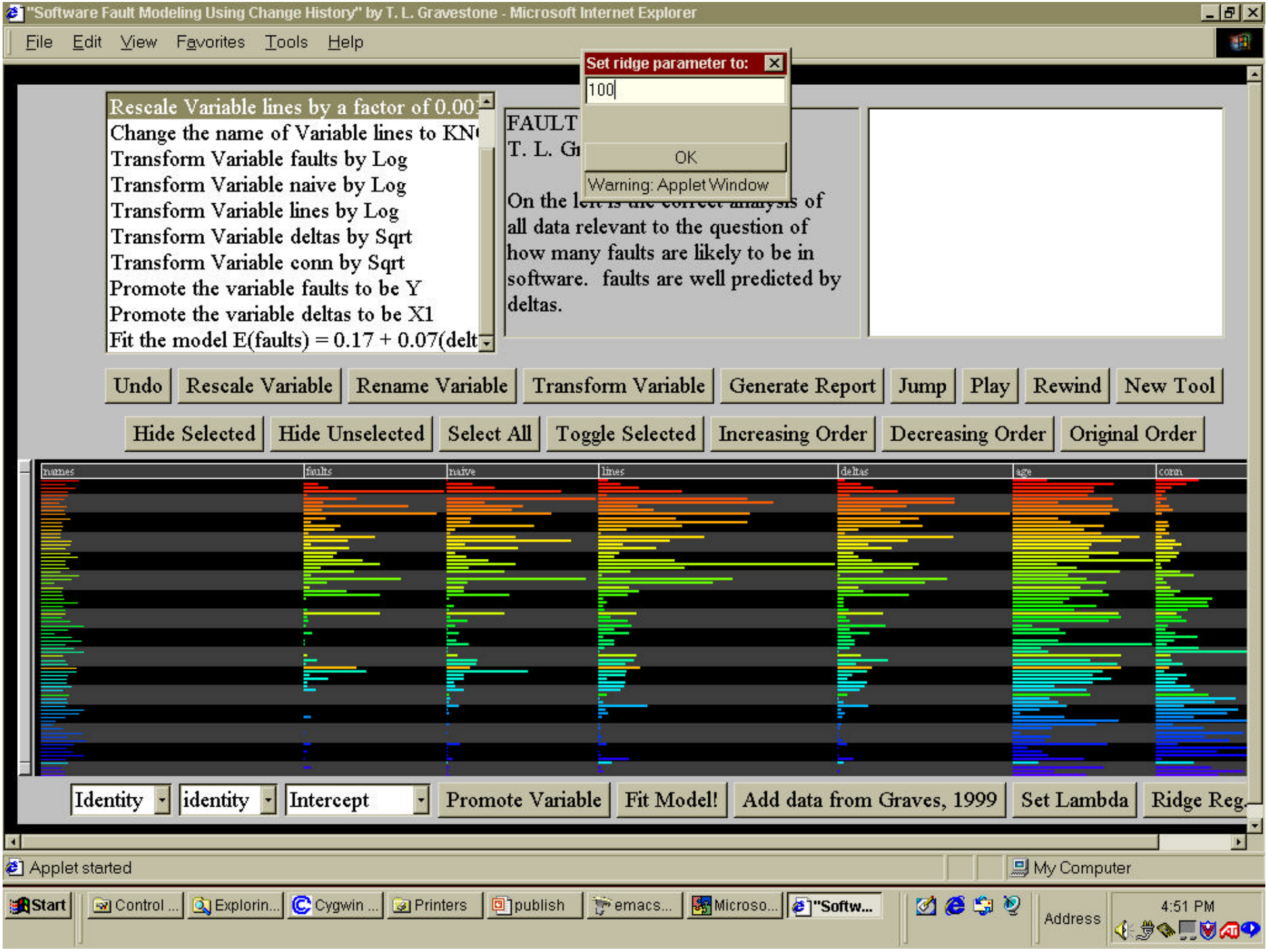
Conclusions

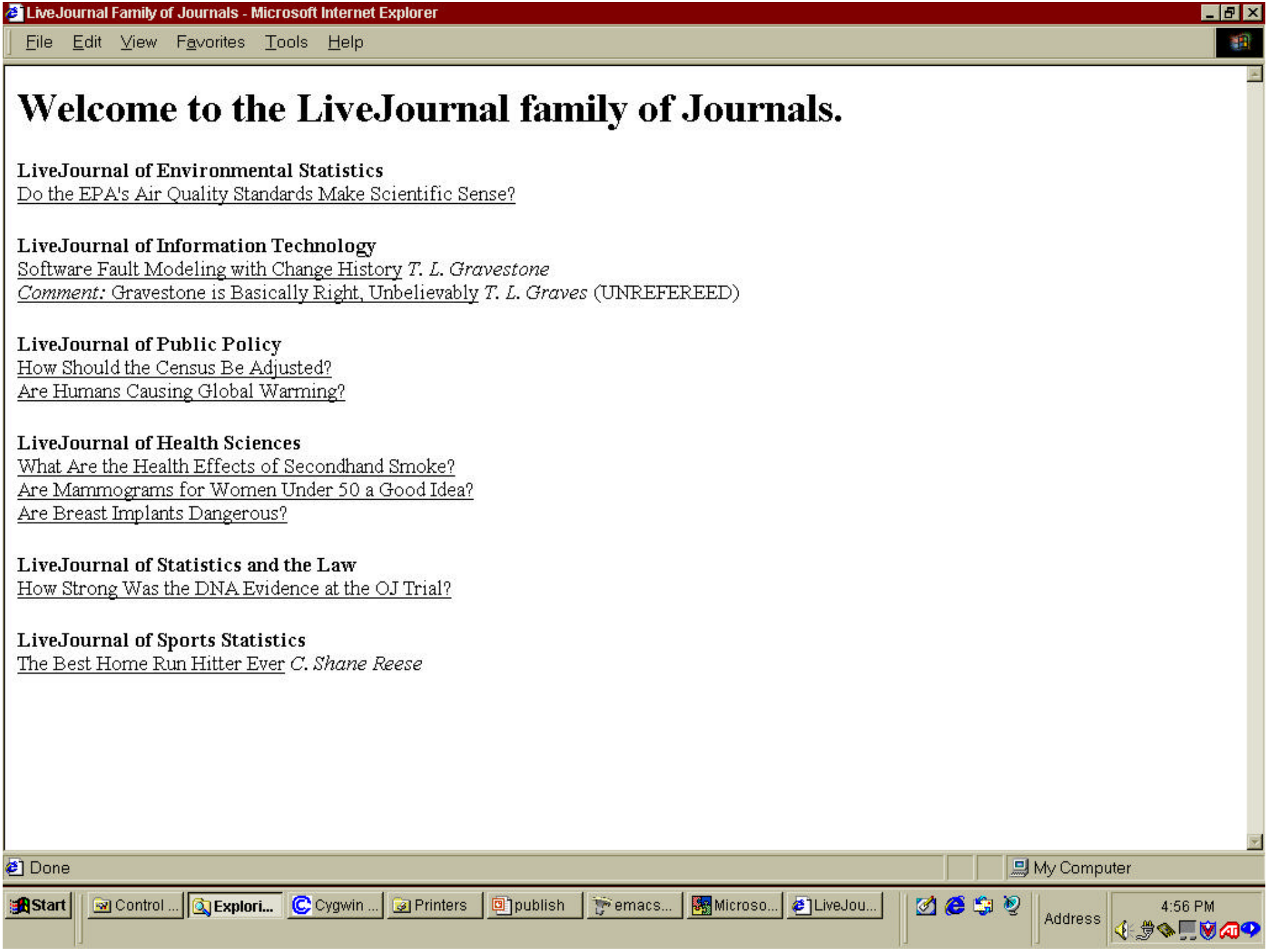
- Applied statistics can benefit from the web in unique ways
- Enriching the journal article reading experience, and reducing barriers to readers' transformation into researchers working on the same problem

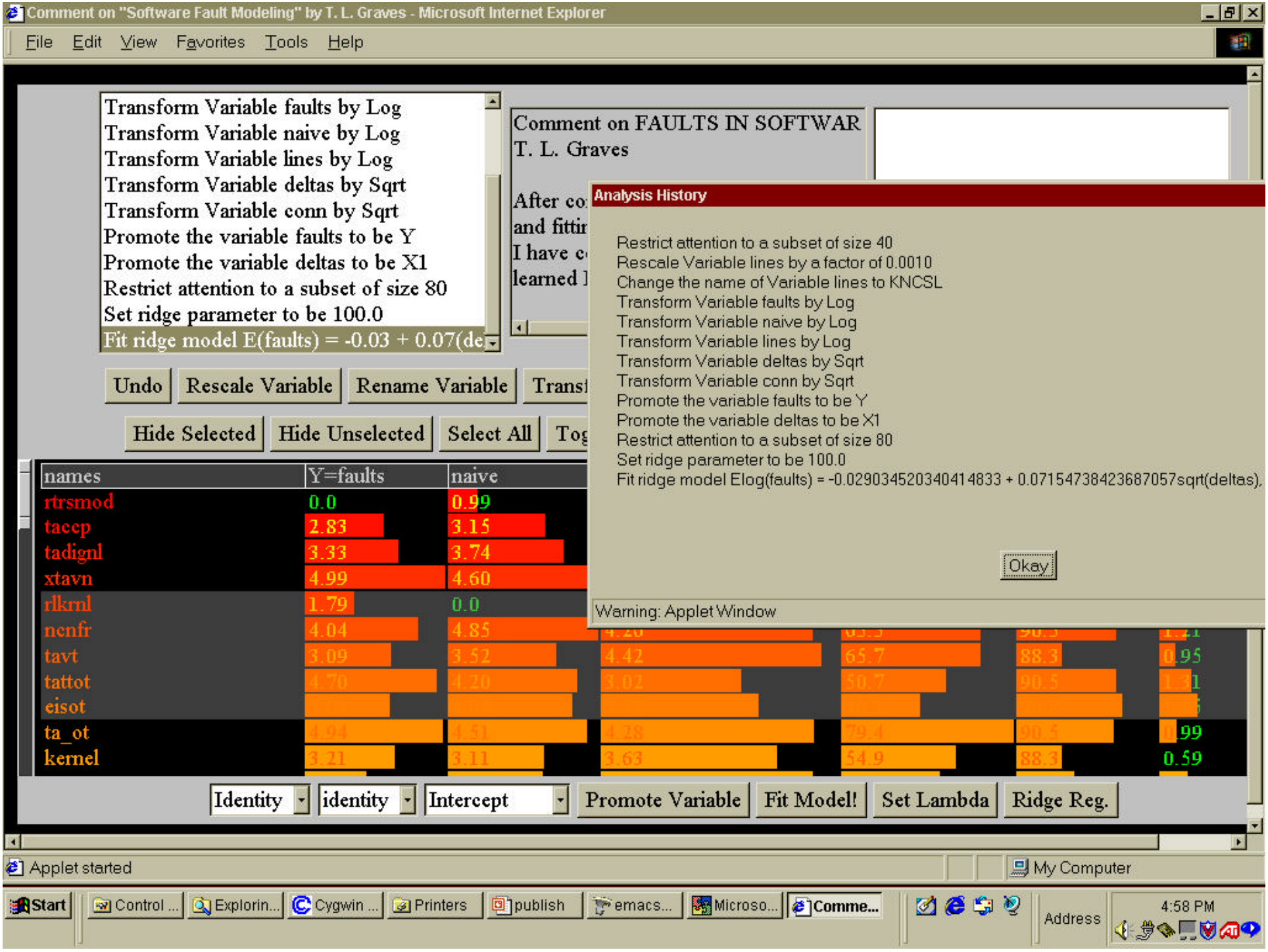

LiveJournal Family of Journals - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail Print Edit Discuss
Address C:\Java\Front.html
Go Link
Welcome to the LiveJournal family of Journals.
LiveJournal of Environmental Statistics
Do the EPA's Air Quality Standards Make Scientific Sense?
LiveJournal of Information Technology
Software Fault Modeling with Change History T. L. Gravestone
LiveJournal of Public Policy
How Should the Census Be Adjusted?
Are Humans Causing Global Warming?
LiveJournal of Health Sciences
What Are the Health Effects of Secondhand Smoke?
Are Mammograms for Women Under 50 a Good Idea?
Are Breast Implants Dangerous?
LiveJournal of Statistics and the Law
How Strong Was the DNA Evidence at the OJ Trial?
LiveJournal of Sports Statistics
The Best Home Run Hitter Ever C. Shane Reese
Done
My Computer
Start
Control ...
Explori...
Cygwin ...
Printers
publish
emacs...
"Softwa...
LiveJou...
4:31
Address charmap


"Software Fault Modeling Using Change History", by T. L. Gravestone - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Rescale Variable lines by a factor of 0.00
Change the name of Variable lines to KN
Transform Variable faults by Log
Transform Variable naive by Log
Transform Variable lines by Log
Transform Variable deltas by Sqrt
Transform Variable conn by Sqrt
Promote the variable faults to be Y
Promote the variable deltas to be X1
Fit the model E(faults) = 0.17 + 0.07(delt
FAULTS IN SOFTWARE
T. L. Gravestone
On the left is the correct analysis of all data relevant to the question of how many faults are likely to be in software. faults are well predicted by deltas.
Undo
Rescale Variable
Rename Variable
Transform Variable
Generate Report
Jump
Play
Rewind
Hide Selected
Hide Unselected
Select All
Toggle Selected
Increasing Order
Decreasing Order
Original Order
names faults naive KNCSL deltas age conn
rtrsmod 0.0 0.99 1.91 840 87.8 4.52
tacep 2.83 3.15 2.50 901 90.6 3.37
tadignl 3.33 3.74 3.04 1343 89.8 1.98
xtavn 4.99 4.60 3.70 2248 89.8 1.36
rlkrnl 1.79 0.0 1.77 506 89.5 1.04
ncnfr 4.04 4.85 4.26 4302 90.5 1.47
tavt 3.09 3.52 4.42 4326 88.3 0.92
tattot 4.70 4.20 3.02 2573 90.5 1.73
eisot 3.04 4.04 3.32 1616 90.8 2.13
ta_ot 4.94 4.51 4.28 6305 90.5 0.99
kernel 3.21 3.11 3.63 3015 88.3 0.35
Identity
identity
Intercept
Promote Variable
Fit Model!
Applet started
My Computer
Start
Control ...
Explori...
Cygwin ...
Printers
publish
emacs...
Microso...
"Softwar...
Address
4:42 PM


"Software Fault Modeling Using Change History", by T. L. Gravestone - Micros
File Edit View Favorites Tools Help
Rescale Variable lines by a factor of 0.00
Change the name of Variable lines to KN
Transform Variable faults by Log
Transform Variable naive by Log
Transform Variable lines by Log
Transform Variable deltas by Sqrt
Transform Variable conn by Sqrt
Promote the variable faults to be Y
Promote the variable deltas to be X1
Fit the model E(faults) = 0.17 + 0.07(delt
FAUL
T. L.
On the
all data
how m
softwa
deltas.
Undo
Rescale Variable
Rename Variable
Hide Selected
Hide Unselected
Select All
Analysis History
Rescale Variable lines by a factor of 0.0010
Change the name of Variable lines to KNCSL
Transform Variable faults by Log
Transform Variable naive by Log
Transform Variable lines by Log
Transform Variable deltas by Sqrt
Transform Variable conn by Sqrt
Promote the variable faults to be Y
Sort data according to decreasing KNCSL
Restrict attention to a subset of size 56
Sort data according to decreasing age
Restrict attention to a subset of size 49
Promote the variable deltas to be X1
Fit the model Elog(faults) = 0.6395221888951119 + 0.05870799777374004sqrt(deltas), s = 0.98
Okay
Warning: Applet Window
names
Y=faults
naive
KNCSL
X1=deltas
age
conn
Identity
identity
Intercept
Promote Variable
Fit Model!
Applet started
My Computer
Start
Control ...
Explorin...
Cygwin ...
Printers
publish
emacs...
Microso...
"Softw...
Address
4:46 PM


"Software Fault Modeling Using Change History" by T. L. Gravestone - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Set ridge parameter to:
100
OK
Warning: Applet Window
Rescale Variable lines by a factor of 0.00
Change the name of Variable lines to KN
Transform Variable faults by Log
Transform Variable naive by Log
Transform Variable lines by Log
Transform Variable deltas by Sqrt
Transform Variable conn by Sqrt
Promote the variable faults to be Y
Promote the variable deltas to be X1
Fit the model E(faults) = 0.17 + 0.07(delt
FAULT
T. L. Gr
On the left is the correct analysis of all data relevant to the question of how many faults are likely to be in software. faults are well predicted by deltas.
Undo
Rescale Variable
Rename Variable
Transform Variable
Generate Report
Jump
Play
Rewind
New Tool
Hide Selected
Hide Unselected
Select All
Toggle Selected
Increasing Order
Decreasing Order
Original Order
names
faults
naive
lines
deltas
age
conn
Identity
identity
Intercept
Promote Variable
Fit Model!
Add data from Graves, 1999
Set Lambda
Ridge Reg.
Applet started
My Computer
Start
Control ...
Explorin...
Cygwin ...
Printers
publish
emacs...
Microso...
"Softw...
Address
4:51 PM


LiveJournal Family of Journals - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Welcome to the LiveJournal family of Journals.
LiveJournal of Environmental Statistics
Do the EPA's Air Quality Standards Make Scientific Sense?
LiveJournal of Information Technology
Software Fault Modeling with Change History T. L. Gravestone
Comment: Gravestone is Basically Right, Unbelievably T. L. Graves (UNREFEREED)
LiveJournal of Public Policy
How Should the Census Be Adjusted?
Are Humans Causing Global Warming?
LiveJournal of Health Sciences
What Are the Health Effects of Secondhand Smoke?
Are Mammograms for Women Under 50 a Good Idea?
Are Breast Implants Dangerous?
LiveJournal of Statistics and the Law
How Strong Was the DNA Evidence at the OJ Trial?
LiveJournal of Sports Statistics
The Best Home Run Hitter Ever C. Shane Reese
Done
My Computer
Start
Control ...
Explori...
Cygwin ...
Printers
publish
emacs...
Microso...
LiveJou...
Address
4:56 PM


Comment on "Software Fault Modeling" by T. L. Graves - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Transform Variable faults by Log
Transform Variable naive by Log
Transform Variable lines by Log
Transform Variable deltas by Sqrt
Transform Variable conn by Sqrt
Promote the variable faults to be Y
Promote the variable deltas to be X1
Restrict attention to a subset of size 80
Set ridge parameter to be 100.0
Fit ridge model E(faults) = -0.03 + 0.07(de
Comment on FAULTS IN SOFTWAR
T. L. Graves
Undo
Rescale Variable
Rename Variable
Hide Selected
Hide Unselected
Select All
Analysis History
Restrict attention to a subset of size 40
Rescale Variable lines by a factor of 0.0010
Change the name of Variable lines to KNCSL
Transform Variable faults by Log
Transform Variable naive by Log
Transform Variable lines by Log
Transform Variable deltas by Sqrt
Transform Variable conn by Sqrt
Promote the variable faults to be Y
Promote the variable deltas to be X1
Restrict attention to a subset of size 80
Set ridge parameter to be 100.0
Fit ridge model Elog(faults) = -0.029034520340414833 + 0.07154738423687057sqrt(deltas),
Okay
Warning: Applet Window
names
Y=faults
naive
rtrsmod 0.0 0.99
tacep 2.83 3.15
tadignl 3.33 3.74
xtavn 4.99 4.60
rlkrnl 1.79 0.0
ncnfr 4.04 4.85
tavt 3.09 3.52 4.42 65.7 88.3 0.95
tattot 4.70 4.20 3.02 50.7 90.5 1.31
eisot
ta_ot 4.94 4.51 4.28 79.4 90.5 0.99
kernel 3.21 3.11 3.63 54.9 88.3 0.59
Identity
identity
Intercept
Promote Variable
Fit Model!
Set Lambda
Ridge Reg.
Applet started
My Computer
Start
Control ...
Explorin...
Cygwin ...
Printers
publish
emacs...
Microso...
Comme...
Address
4:58 PM